HLA Imputation Using Transformer Model

Divyansh Yadav
Dept. of Computer Science
Bowling Green State University
Bowling Green, OH, USA
dyadav@bgsu.edu

***Abstract*—In this paper, we investigate the application of Transformer-based architectures for sequence processing tasks. Transformers have shown remarkable success in various natural language processing tasks due to their attention mechanisms and parallelizability. We aim to understand how the Transformer model performs, especially in complex sequence tasks like machine translation. The self-attention mechanism in Transformers allows for capturing global dependencies in sequences effectively. Through experiments conducted on HLA genotype sequences, we evaluate the Transformer model's performance in terms of metrics such as BLEU score, contextual sensitivity, and computational efficiency. Our results indicate that the Transformer model surpasses traditional approaches in processing sequences, demonstrating its efficacy and potential for enhancing sequence understanding across various domains.**

***Index Terms*—HLA, Transformer, NLP**

I. Introduction

Human Leukocyte Antigens (HLA) proteins are like ID tags on the surface of our cells, helping the immune system tell apart what's naturally part of the body and what's an outside invader, like viruses or bacteria. These proteins come from a wide range of genes, making them very diverse. This diversity is key when it comes to things like organ transplants, where matching these proteins between donors and recipients helps prevent the body from rejecting the new organ. Also, differences in these HLA genes can affect how likely someone is to develop autoimmune diseases, where the body mistakenly attacks its own cells. This makes understanding HLA genes super important for doctors and researchers.

HLA imputation is a fancy way of guessing someone's HLA gene setup based on other genetic information they have, like SNPs (small changes in DNA). It's a handy workaround to the more expensive and complicated direct testing of HLA genes. This guessing game is incredibly useful for studying diseases related to the immune system and for finding matching organ donors. Getting these guesses right is super important for personalized medicine, where treatments are tailored to an individual's genetic makeup, and for making organ transplants more successful. This progress in figuring out how our genes affect our health and how we respond to treatments is really exciting for improving medical care and making sure organ transplants go smoothly.

The Transformer machine learning model holds immense importance in bioinformatics due to its exceptional ability to analyze and interpret complex biological data. Through its innovative self-attention mechanisms, Transformers excel at capturing intricate patterns and dependencies in biological sequences, such as DNA, RNA, and protein sequences. This capability facilitates various crucial tasks in molecular biology, including sequence analysis, genomic variation analysis, and prediction of genotype-phenotype associations. Additionally, Transformer models play a vital role in drug discovery and development by predicting molecular interactions and identifying potential drug candidates. As the field of bioinformatics continues to generate vast amounts of data, the Transformer model's capacity to handle large-scale datasets and extract meaningful insights is paramount for advancing our understanding of biology, disease mechanisms, and therapeutic interventions.

We Utilized machine learning to forecast high-resolution HLA genotypes using a transformer model. This method improves the precision and efficacy of HLA-type identification, supporting disease association research and tailored medication. It is essential to medical research and transplantation.

II. LITERATURE REVIEW

In the article "Role of Human Leukocyte Antigens (HLA) in Autoimmune Diseases" [1] talks about how HLA genotyping plays a significant role in the diagnostic work-up of autoimmune diseases. It is routinely performed as part of the diagnostic process for certain autoimmune diseases. HLA genotyping is essential in determining donor-recipient immune compatibility in organ transplantation. Additionally, HLA genotyping is used to identify genetic susceptibility to autoimmune diseases such as celiac disease, rheumatoid arthritis, spondyloarthritis, and Behc¸et's disease. The presence of specific HLA alleles, such as HLA-DQ2/DQ8 in celiac disease and HLA-B27 in spondyloarthritis, can aid in the diagnosis of these conditions. Furthermore, HLA genotyping can help exclude autoimmune diseases in high-risk individuals and reduce the need for invasive diagnostic procedures. Overall, HLA genotyping has become an integral part of the diagnostic work-up for autoimmune diseases, providing valuable information for patient management and treatment decisions. The paper, titled "A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes" [2] by Tatsuhiko Naito et al., presents DEEPHLA, a deep learning-based method designed to impute HLA genotypes accurately, especially focusing on the challenges posed by the

diversity in allele frequency across different ethnic groups. Traditional HLA imputation methods struggle with infrequent alleles, leading to reduced reliability in trans-ethnic major histocompatibility complex (MHC) fine-mapping due to allele frequency spectrum heterogeneity among different ethnicities. DEEPHLA addresses these issues by leveraging deep learning techniques, achieving significantly higher accuracy in imputing low-frequency and rare alleles compared to existing methods.

The paper "Attention Is All You Need," by Vaswani et al.[3], introduces the Transformer model, a novel approach to machine translation that departs from the traditional sequence-to-sequence models reliant on recurrent neural networks (RNNs) or convolutional neural networks (CNNs). Instead, the Transformer leverages a mechanism called "attention" to process data in parallel, significantly improving translation quality and training speed. Applications of transformer-based language models in bioinformatics [4] talks about significant contributions to bioinformatics research in several ways. These models, such as the vanilla transformer, BERT, and GPT-3, have demonstrated remarkable interpretability and adaptability, leading to their application in various bioinformatics tasks like

De novo drug generation: A model based on transformer architecture has been proposed to generate realistic lead compounds using only the amino acid sequence information of the target protein (Grechishnikova, 2021).

Molecular generation: Transformer-based models have been used for molecular generation, leveraging masking self-attention mechanisms to capture long-range dependencies in order to generate new molecules (Bagal et al., 2022). Basic Sequence Analysis: Transformer-based language models have been utilized for tasks such as DNA/RNA sequence analysis, gene expression, and proteomics. They have shown promise in analyzing and interpreting biological sequences. This shows the capabilities of transformer models in bioinformatics which predict the output with great accuracy and can work for our complex HLA structure

Another paper Transformer- based on deep learning for predicting protein properties in the life sciences[5] talks about Transformer-based deep learning models for predicting protein properties in the life sciences. It notes the rapid advancement of natural language processing models, particularly the Transformer model, in this area. The review emphasizes the potential of deep learning techniques, especially those from natural language processing, to uncover intricate patterns in protein data for property prediction. It also includes a comparative analysis between traditional protein features and those extracted by Transformer models, illustrating the latter's promise as versatile feature extractors. Ultimately, the review anticipates that Transformer-like models will likely become the standard approach for computational biology and bioinformatics tasks shortly.

A. Limitations Of Previous Studies

The dependence of machine learning models for training on large and diverse datasets may be a major disadvantage. The model's predictions might be biased if these datasets aren't diverse or thorough, which means the results will vary according to demographics and ethnicity. also, the deep neural networks are mostly black box and sometimes it's hard to interpret the results.

III. Proposed approach

The transformer model with its extremely effective and efficient architecture is a ground-breaking development in the field of natural language processing (NLP) for a variety of language-related activities. Transformers were first presented by Vaswani et al. in their 2017 paper "Attention is All You Need" [2] and they have since grown to be an essential component of contemporary NLP systems, outperforming convolutional neural networks (CNNs) and recurrent neural networks (RNNs) in many applications. The transformer is based on the principle of self-attention, which allows the model to process individual words in a phrase and determine their relative importance. Transformers can evaluate every word in a sentence at once, unlike traditional sequential models like RNNs, which analyze words one after the other. This makes transformers extremely effective in capturing long-range dependencies and contextual information.

The representation of input tokens (words or subwords) as vectors is the first crucial step in the process by which self-attention works. Then, using learned linear transformations, these vector representations are divided into three sets of vectors: Query, Key, and Value. The key to self-attention is calculating similarity scores between the Key vectors of all other tokens and the Query vector of each token. The outcome of this calculation is a set of scores that indicate how much emphasis or attention each token ought to provide to other tokens. These scores are run through a softmax function to provide attention weights, which normalize the data and make it comprehensible as a probability. Each token in the sequence should receive a different amount of attention based on these weights. The Value vectors are then added together to get a weighted total by utilizing the attention weights that were previously acquired. Each token's output representation is this weighted sum, which is subsequently supplied into the transformer architecture's further stages.

Through the self-attention process, transformers are able to extract rich contextual information from the complete input sequence, which makes it easier to comprehend the deep links between words and helps the model learn new patterns. Furthermore, transformers may be effectively trained on big datasets with the use of contemporary hardware accelerators like GPUs and TPUs thanks to the parallel nature of self-attention. Transformers have thus far shown impressive performance in a variety of natural language processing (NLP) applications, such as sentiment analysis, text production, and language translation. Transformers continue to propel developments in natural language processing (NLP) and reshape the artificial intelligence landscape with their capacity to manage distant dependencies and grasp subtle semantic links.

Fig1 shows the transformer model architecture that was proposed in Attention Is All You Need [3].

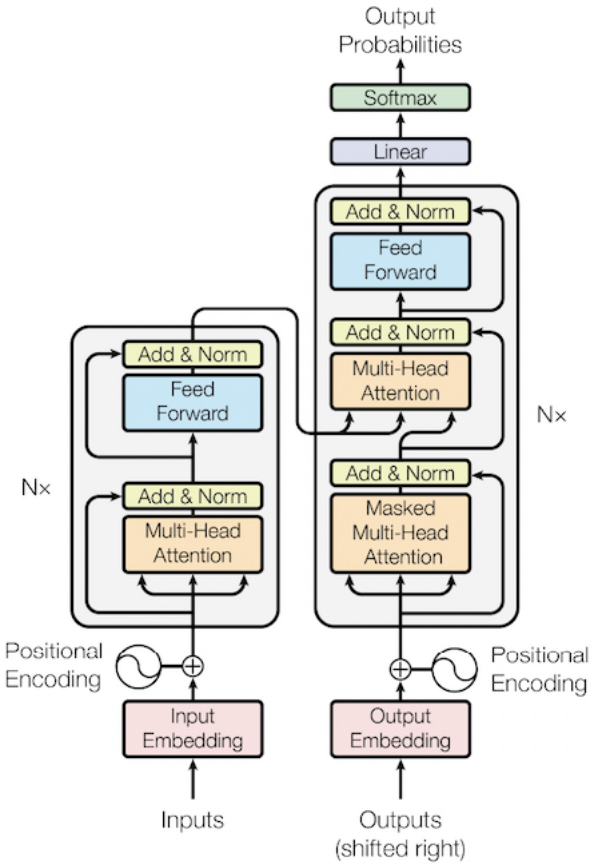

Fig. 1. The Transformer - model architecture

In our case with 1095 items, the dataset provides a list of genotypes for the human leukocyte antigen (HLA), a crucial aspect of immunogenetics. Every dataset item is organized using two sets of HLA types: one set in a high-resolution or more detailed format (e.g., A*03:01, A*32:01), and the other set in a simplified or low-resolution style (e.g. A3, A24). Low-resolution versions provide a wide categorization of HLA alleles, which are frequently used in clinical settings for preliminary immune response studies. On the other hand, high-resolution variants offer more detailed genetic data, which is essential for accurate medical uses such as organ transplant matching and comprehensive immunological studies.

There are multiple HLA genes, and A, B, and C are among the major ones in class I. MHC class I molecules present fragments of proteins (peptides) from inside the cell to T cells. HLA genes, renowned for their extensive polymorphism, play a crucial role in regulating immune responses and are associated with a range of conditions such as autoimmune diseases, infectious diseases, and transplant compatibility. Conventional methods for HLA typing are frequently expensive and time-intensive, underscoring the demand for more streamlined computational methodologies. Leveraging the capacity to process vast datasets and unveil intricate patterns, machine learning presents a robust alternative for addressing this challenge.

Machine learning techniques based on transformer algorithms have been utilized for predicting HLA genotypes using genomic data, including single nucleotide polymorphisms (SNPs), we used encoder and decoder RNN with self-attention to encode the vector, the start of the sentence and end of the sentence was added and we split the data into train and test split with 80:20 ratio between train and test, then the data was loaded in the batches of 32 and trained with 128 hidden layers after the training process we have used the test data to test the accuracy in form of BLEU score.

IV. MODEL EVALUATION AND RESULTS

In this section, we will discuss the results and approach used for evaluating those results

A. MODEL EVALUATION

The BLEU score (Bilingual Evaluation Understudy) is a metric used to evaluate the quality of text that has been machine-translated from one language to another. The BLEU score measures how closely the machine-generated translations resemble a set of high-quality human translations.

The score ranges from 0 to 1, where 1 means a perfect match with the reference translation(s). However, BLEU scores are often expressed as a percentage (0 to 100). Higher BLEU scores indicate translations that are closer to human quality, suggesting that the translated text is more accurate and fluent.

B. RESULTS

In this experiment, we have utilized the transformer model which relies on attention mechanisms, sidelining traditional recurrent or convolutional layers. Its architecture, composed of encoders and decoders, excels in understanding the context and relationships in text through self-attention, allowing it to weigh the importance of each word in a sentence. This design enables significant parallel processing, making training on large datasets much more efficient. Widely used for tasks like translation and text summarization. Our problem with HLA translation fits in the domain of language translation and transformer models have shown better accuracy in text translation. We had data samples of 1095 samples, having low- and high-resolution HLA genotypes, we started by reading the data and splitting the data into 80-20 ratios for train and test split. Once the data split data is ready we need to encode the data before training, so we have used an EncoderRNN which is a recurrent neural network used to encode the text also for decoding we have AttnDecoderRNN which is a recurrent neural network with self-attention, we used this encoder and decoder and train the model using the data. For assessing the performance of our models, we applied the BLEU score metric, a well-regarded standard in machine translation evaluations. This metric quantifies the similarity between the text produced by the Transformer models and the reference texts. A higher BLEU score signifies that the generated sequences closely match human-annotated, high-quality HLA types, showcasing the Transformer models' capability to precisely predict HLA types. On checking the BLEU score the score was around 1 which signifies a better prediction.

V. Conclusion and Discussion

Enhanced typing resolution holds paramount importance across a spectrum of medical and research domains, notably in organ transplantation, disease predisposition assessment, and tailored medical interventions. Consequently, ensuring high-resolution Human Leukocyte Antigen (HLA) typing is imperative for effective organ transplantation endeavors. However, cost considerations often prompt the utilization of lower resolution typing. Leveraging machine learning techniques to translate lower resolution HLA data into higher resolution formats can significantly aid clinicians in improving diagnostic accuracy.

In this study, we employed a dataset comprising 1095 entries of both low and high-resolution HLA records, partitioned into an 80:20 ratio for training and testing respectively. Subsequently, the data underwent preparation before training a Transformer model. The Transformer, a machine learning architecture renowned for its proficiency in text comprehension and processing, operates distinctively from conventional models by concurrently assimilating the entire text and employing self-attention mechanisms to discern salient features.

Functioning akin to an adept reader discerning pivotal aspects of a narrative in a single glance, the Transformer comprises an encoder and decoder components, facilitating breakdown and utilization of input text to generate responses or translations. Its capacity to preserve word sequence and capture long-range dependencies within text has revolutionized tasks such as translation and summarization.

Our methodology entailed employing an encoder-decoder framework integrated with self-attention mechanisms for data interpretation and training. Subsequently, the trained model was evaluated using the BLEU (Bilingual Evaluation Understudy) score, a metric commonly used in assessing the fidelity of machine-generated translations, particularly in the realm of translation tasks. In our study, achieving a BLEU score of 0.65 indicates precise prediction of high-resolution HLA values, thereby affirming the model's accuracy.

VI. limitation

Transformer models have indeed transformed the landscape of Natural Language Processing (NLP) with their impressive performance, but they do come with their fair share of drawbacks. Their hefty computational requirements can pose accessibility challenges, while their substantial memory needs limit their application in settings with limited resources. Although they excel at capturing long-distance relationships, Transformers struggle with tasks demanding a deep understanding of extensive context due to their restricted contextual window. Furthermore, their black box workings present hurdles in understanding their decisions, especially in fields like law or medicine where transparency is vital.

The success of the transformer model depends a lot on the data it learns from and generally, they need a huge amount of data for learning. In this study, the type and amount of data, with 1095 rows, really matter. If the data doesn't represent a wide range of situations or doesn't have enough variety, the machine might learn too much from it and not be able to adapt to new situations very well. This is especially tricky in genetics because understanding and predicting genetic stuff need a lot of different kinds of data. While we often use BLEU ratings to see how well machine translation works, they might not be the best fit for genetics. BLEU ratings mainly look at how similar the predicted and real sequences are, but they don't think about how important those predictions are in the genetic world. In genetics, our evaluation methods must consider not just how similar sequences are, but also how biologically meaningful the predictions are for helping us understand and make decisions about genes and diseases.

In the context of Human Leukocyte Antigen (HLA) and genetics, diversity, and nationality play crucial roles in understanding immune responses, disease susceptibility, and transplantation compatibility. HLA molecules are essential components of the immune system, responsible for presenting foreign antigens to immune cells, thus triggering immune responses. The HLA system is highly polymorphic, meaning that there are numerous variations or alleles within the population.

Diversity in HLA alleles across different populations is significant because it reflects the genetic variability among individuals. Different populations have distinct genetic backgrounds shaped by evolutionary forces, migration patterns, and demographic history. As a result, the distribution of HLA alleles varies among populations.

VII. Future Work

Broadening the scope of data collection to encompass diverse geographic regions contributes significantly to enhancing the richness and representativeness of the dataset. Given the dependency of Human Leukocyte Antigen (HLA) data on ethnicity and demographic factors, augmenting the dataset with samples from various global locales fosters greater inclusivity and ensures a more comprehensive understanding of HLA variations across populations.

By incorporating data from diverse ethnic backgrounds and geographic origins, the dataset becomes more robust, enabling the model to capture a broader spectrum of HLA diversity. This, in turn, enhances the model's ability to generalize and adapt to variations in HLA profiles encountered in clinical settings worldwide.

Furthermore, refining the model's performance can be achieved through hyperparameter tuning or even customizing the architecture, such as designing a bespoke transformer model tailored specifically for HLA typing tasks. Fine-tuning the model's parameters and architecture empowers it to discern subtle nuances in HLA data and optimize predictive accuracy, thereby advancing its utility in clinical applications.

VIII. ACKNOWLEDGEMENT

I extend my sincere gratitude to Dr. Rob Green for his invaluable guidance, expertise, and provision of computational resources, which have been significant in facilitating the completion and success of this project.

REFERENCES

[1] Role of Human Leukocyte Antigens (HLA) in Autoimmune Diseases by Gergely Bodis, Victoria Toth, Andreas Schwarting

[2] A deep learning method for HLA imputation and trans-ethnic MHC fine-mapping of type 1 diabetes by Naito T, Suzuki K, Hirata J, Kamatani Y, Matsuda K, Toda T, Okada Y.

[3] Attention Is All You Need by Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin

[4] Applications of transformer-based language models in bioinformatics: a survey Shuang Zhang, Rui Fan1, Yuti Liu, Shuang Chen, Qiao Liu and Wanwen Zeng

[5] Transformer- based deep learning for predicting protein properties in the life sciences Abel Chandra , Laura Tünnermann , Tommy Löfstedt , Regina Gratz